Non-genetic cell-to-cell variability: theory and experiments

Schwabe, A.

2014

document version
Publisher's PDF, also known as Version of record

Chapter 3

Volume scaling of the exact mRNA concentration indicates homeostasis and explains cell-to-cell heterogeneity

Contents

in collaboration with H. Kempe, F. Crémazy, P.J. Verschure, and F.J. Bruggeman
Manuscript submitted
Author contributions: Mannus Kempe performed the experiments, while I focused on the theoretical parts of the paper. Mannus and I wrote the image analysis scripts together.

Cell-to-cell variation in gene expression can be measured by counting the number of mRNA molecules per cell. However, reaction rates in living cells are determined by concentrations rather than molecule numbers. We combined single-molecule mRNA counting with single-cell volume measurements to determine the statistics of mRNA concentrations in human cells. We studied three cell lines that differ only in the genomic integration site of an identical constitutively expressed reporter gene. We observe gene-location dependent transcription activity. The mRNA number per cell varies proportional with the cell volume in all three cell lines, indicating concentration homeostasis. We observe and theoretically explain the reduction in gene expression variability, when expressed in mRNA concentration units rather than in mRNA numbers. The functional noise in mRNA concentration is predominantly determined by gene expression noise. This study highlights that the coupling between stochastic gene-expression dynamics and cell-volume growth sets the functional noise of living cells.

3.1 Introduction

Spontaneous fluctuations in the activities of molecular processes cause heterogeneity in the molecular composition of isogenic cells [Ozbudak 2002, Elowitz 2002, Sigal 2006]. Cell-to-cell variability has been observed in mRNA [Golding 2005, Raj 2006, Zenklusen 2008] and protein levels [Ozbudak 2002, Elowitz 2002, Yu 2006, Sigal 2006], in the timing of molecular processes [Amir 2007, Di Talia 2007], and in cellular growth rates [Boulineau 2013]. The causes of molecular noise involve molecules occurring at low numbers per cell, such as transcription factors or mRNAs, that tend to show large, spontaneous deviations relative to their mean number within the cell population [Paulsson 2004]. These deviations (fluctuations) can be caused by cell division [Huh 2011b], transcription bursting [Suter 2011], or transient imbalances between molecular synthesis and degradation rates that occur spontaneously through thermal noise ('intrinsic noise') or due to fluctuations in the number of regulators ('extrinsic' noise) [Thattai 2001, Elowitz 2002, Swain 2002, Paulsson 2004]. Extrinsic noise indicates that fluctuations can propagate through the entire molecular network of a cell [Pedraza 2008]. As a result, the molecular compositions of cells can be highly variable and cause heterogeneity in differentiation decisions [Wernet 2006], stress response magnitudes [Veening 2005], and the survival prospects of cells after drug exposure [Spencer 2009].

Early studies on stochastic gene expression relied on fluorescent proteins to assess protein noise by either taking snapshots [Ozbudak 2002, Elowitz 2002] or by real-time fluorescence imaging [Rosenfeld 2005, Sigal 2006]. More recently, single-molecule mRNA counting has been introduced [Golding 2005, Raj 2006, Raj 2008a] as a method for absolute quantification of mRNA numbers. The advantage of single-transcript counting with single-molecule RNA FISH (smFISH) [Raj 2006, Raj 2008a, Zenklusen 2008, Youk 2010] is that it does not require genetic

engineering. Specific DNA-probes tagged with fluorescent dyes are used to visualize individual mRNA molecules within fixed cells (fig. 3.1A).

From a cell-biological perspective, the statistics of absolute numbers of a molecule per cell, as obtained with smFISH, do not always reflect heterogeneity that causes changes in the rates of molecular processes. Since rates of molecular processes depend on molecule concentrations, a cell with twice the volume and an equal copy number of a molecule compared with another cell, will have half the rate of processes involving that molecule. Concentration noise can therefore be seen as functional noise and this is more informative about cell-to-cell variability than noise in the number of molecules per cell. However, single-cell concentration values for mRNA data have so far not been reported. Here we report on the functional noise in mRNA levels.

The mRNA concentration in a single cell equals the number of mRNA molecules in this cell divided by its volume. The statistics of mRNA concentrations in single cells therefore results from the relation between the volume distribution and the mRNA copy number distribution across a population of cells. The latter results from the stochasticity of gene expression and mRNA decay [Paulsson 2004] and varies amongst mRNAs [Bar-Even 2006, Taniguchi 2010] whereas the former depends on cellular growth [Tzur 2009] and impacts all transcript concentrations equally. At steady-state growth of the cell population, which is usually the condition for single-cell studies [Sigal 2006, Rosenfeld 2005], every daughter cell doubles on average its volume and mRNA numbers while it ages to become a mother cell ready for division. Therefore, part of the cell-to-cell variability in the mRNA number per cell derives from the cell-age dependency of the mRNA numbers [Marathe 2010]. The transcript concentration compensates for this cell-age effect to an extent that depends on the scaling of the mean mRNA number with cell volume. The cell-age dependency on concentration noise becomes negligible when the cell volume increases as function of the cell age at an equal rate as the mRNA number as function of cell age. A discrepancy between these rates will set the cell-age dependent contribution to concentration noise. This discrepancy depends on how the gene activity scales with cell age, e.g. via replication or cell-cycle dependent regulation [Farkash-Amar 2012], and how the cell volume increases with age, e.g. linearly, biphasically, or exponentially [Cohen 2009, Sigal 2006, Rosenfeld 2005, Cookson 2010]. Understanding of mRNA concentration noise requires in addition the determination of how the variance in the mRNA numbers scales with cell volume.

In this study, we measured single cell mRNA concentrations by quantifying the volume of single cells and their mRNA numbers using confocal microscopy. We studied three human cell line variants that express the same constitutively-expressed gene from a different genomic location to identify gene-location dependent effects. In order to attain robust statistics of the volume dependency of the mRNA number statistics, we studied nearly a 1000 single cells of each cell line. As a constitutively-expressed gene is representative for a large class of human genes, this study improves our general understanding of how cell-to-cell variability of transcript concentrations is determined in single human cells.

3.2 Results

3.2.1 Single-cell transcript data indicates gene-location dependent mRNA expression

We analyzed three cell lines derived from the same human cell line (HEK293) (described in [Gierman 2007]). Each cell line has a single random insertion of the same GFP reporter construct controlled by a constitutive phosphoglucokinase (PGK) promoter [Gierman 2007] (section 3.3.1.1). We determined the statistics of the GFP mRNA levels in single cells with smFISH (fig. 3.1A). The probe set contained 35 probes of 17 to 18 nucleotides coupled with a fluorescent label (section 3.3.7). Images of single cells were obtained with confocal microscopy on smFISH treated cells counterstained with DAPI. Individual cells were recorded as 52 z-stack images (300 nm/slice; fig. 3.1D). Lamina staining confirmed that the DAPI staining correctly identifies the nuclear envelope (fig. 3.7). Based on the co-localization with the DAPI signal, we assigned a mRNA molecule to be either nuclear or cytoplasmic. An overview of the transcript statistics is given in fig. 3.1B and the mRNA distributions are shown in fig. 3.1E-G and fig. 3.14.

For cell line I, the number of mRNA molecules expressed per cell (m) was on average 45.7 mRNA, obtained from a dataset containing 838 cells. The coefficient of variation (cv; std/mean) indicates that the standard deviation is about 45% of the mean. Approximately 72% (Δ^{σ}) of the cells have mRNA numbers that deviate less than one standard deviation (σ) from the mean mRNA number (μ). The symmetry of the mRNA number distribution is indicated by the 13.8% of cells which have mRNA numbers below 25 ($\mu - \sigma$) and the 13.9% that have higher than 66 mRNA molecules ($\mu + \sigma$) (fig. 3.14).

Colocalization of a mRNA spot with the DAPI signal enabled us to calculate the mRNA number in the nucleus of these cells. The mean number of mRNA per nucleus (m_N) was 9.7 molecules with a standard deviation of 5.4. The cytoplasmic mRNA number (m_C) follows directly from $m - m_N$ and is 36.0 mRNA molecules. The number of mRNA transcripts appears to be lower in the nucleus than in the cytoplasm, indicating that the lifetime of the mRNA in the cytoplasm is higher than the residence time in the nucleus. Compared to the cv of the mRNA in the nucleus, which is 56%, the cytoplasmic mRNA numbers are less noisy with a standard deviation of about 48% of the mean. This higher noise in nuclear mRNA numbers is mostly explained by the higher intrinsic noise contribution ($1/\mu_N$) in the nucleus. The fact that $(\sigma/\mu)^2$ exceeds $1/\mu$ for nuclear mRNA indicates that part of the gene expression noise derives from extrinsic gene-expression noise [Paulsson 2004]. The correlation coefficient (ρ) between the nuclear and cytoplasmic mRNA numbers per cell indicates that $\approx 36\%$ ($\rho^2 \times 100$) of the variance in cytoplasmic mRNA numbers is explained by the variance in nuclear mRNA numbers.

The same analysis as above can be done for the other two cell lines and yields similar results (fig. 3.1B and fig. 3.15). This data enables us to make a comparison between the different genomic integration sites. The mean expression level between

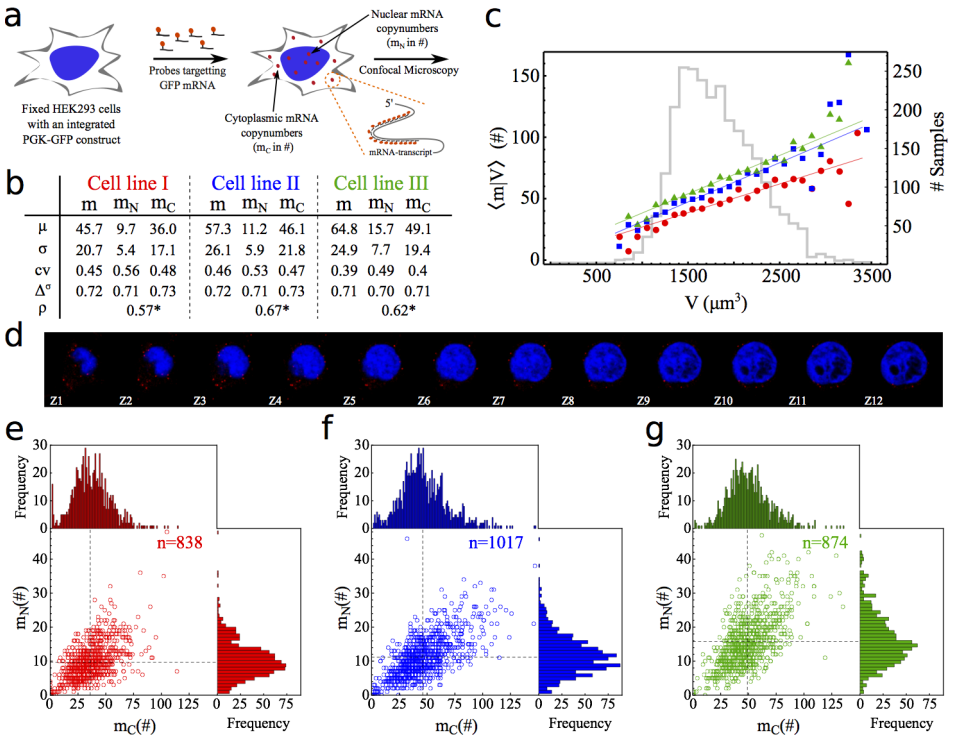

Figure 3.1: **Statistics of single cell mRNA copy number data.** (a) Schematic overview of the smFISH method applied to our reporter gene mRNA. Colocalization of the mRNA molecules with DAPI counter staining identified spots as nuclear mRNA (m_N), others are cytoplasmic mRNA (m_C). (b) Statistics of the mRNA molecules in the cell (m), nucleus (m_N) and cytoplasm (m_C) for the three different cell lines (color coded). Notation: μ=mean, σ=standard deviation, cv= coefficient of variation, Δ^σ= the fraction of samples between $\mu - \sigma$ and $\mu + \sigma$, and ρ= correlation between m_C and m_N, and $^*p < 0.001$ ($H_0 : \rho = 0$). (c) For a specific cell volume (V), the mean mRNA copy number is calculated from the data. This conditional mean ($\langle m|V \rangle$) shows a linear scaling with respect to volume indicating homeostasis in mRNA concentration. The gray histogram in the background shows the total number of cells per volume bin for all three cell lines (bin size = 100 μm^3). Higher counts indicate higher reliability of the corresponding determination of ($\langle m|V \rangle$). A least-squares linear fit is shown for all three cell lines. The explained fraction of the variance in $\langle m|V \rangle$ with this fit is 0.80, 0.77 and 0.84 for cell line I , II and III respectively. (d) Representative confocal images of a cell, with Z1 to Z12 corresponding to subsequent optical sections (z-slices) of the cell. The mRNA molecules are shown in red and the DAPI stained nucleus is shown in blue. (e-g) Scatterplots of m_C and m_N for the three different cell lines. Marginal histograms show the distribution of m_C (top) and m_N (right). The measured number of cells is given by n.

the three cell lines differ (ANOVA; $p < 0.0001$, Supplementary fig. 3.5) on average by 20% and maximally by 40%. These numbers correlate with the protein expression data of the cell lines [Gierman 2007] (Supplementary fig. 3.10). The cell-to-cell

variability in mRNA numbers per cell, measured as the cv, is significantly different (ANOVA; $p < 0.03$, Supplementary fig. 3.15) between the three cell lines. Since the three cell lines differ only in their genomic location of the reporter gene, these differences demonstrate the influence of gene location on expression stochasticity. Additionally, an increase in the mean expression level (μ) does not necessarily cause an increase in the standard deviation (σ) when comparing the different cell lines. This indicates that the mean expression level and the variability are regulated independently, which was previously observed for protein expression data [Viñuelas 2012].

3.2.2 Volume statistics of single cells

In order to assess functional mRNA noise we convert the mRNA numbers to concentration units. The same confocal z-stack images (fig. 3.1D) used for smFISH were used to determine the whole-cell, cytoplasmic and nuclear volumes of the cells by tracing the contours of these compartments (fig. 3.2A). This allows us to obtain mRNA number, volume, and concentration data for each cell.

The measured volume distributions for the whole cell (fig. 3.6) as well as the cytoplasm and nucleus (fig. 3.2C) are positively skewed, which means that there are, relative to the mean, more small than large cells. Similarly shaped distributions have previously been reported for stationary growing cell populations [Tzur 2009]. These positively skewed distributions are due to the formation of two (smaller) daughter cells from each (large) mother cell. The cell volume distributions can be well approximated by theoretical cell volume distributions derived from balanced, exponential growth of the cells (fig. 3.9).

The obtained volume measurements are summarized in fig. 3.2B. The mean cell volume (V) for cell line I is 1800 μm^3, with on average a larger nucleus (979 μm^3) than cytoplasm (822 μm^3). For the volume distribution of cell line I, 70% of the cells have a volume between 1344 μm^3 ($\mu - \sigma$) and 2256 ($\mu + \sigma$) μm^3, 14% are below 1344 μm^3 and the remaining 16% is above 2256 μm^3. Part of this spread is due to the cell-cycle dependency of the volume. The correlation coefficients between the nuclear and the cytoplasmic volume indicate a weak but significant, positive correlation indicating growth of both the nuclear and cytoplasmic volume during cell maturation.

Since the cell lines are isogenic except the integration site of the construct, the measured volume statistics are expected to be similar. Fig. 3.2B and fig. 3.6 confirms this expectation for the three measured volume distributions.

3.2.3 mRNA concentration statistics of single cells indicate mRNA concentration homeostasis

Next, we combined the mRNA number and volume data of each cell to determine the statistics of cellular, cytoplasmic, and nuclear mRNA concentrations (fig. 3.3). Figure 3.3B shows that the mean mRNA concentration differs between the three cell

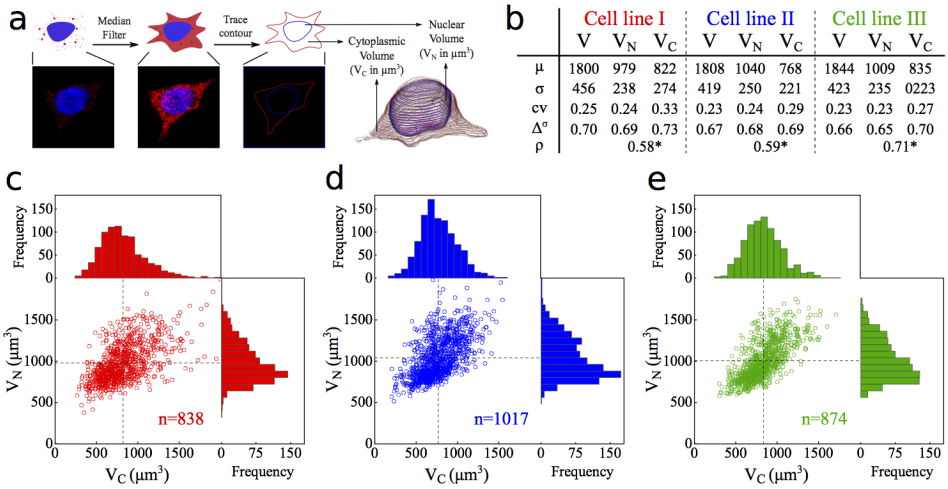

Figure 3.2: **Statistics of single cell volume data.** (a) Overview of the determination of the cell volumes. The background intensity was used to track the contour of the cell and the DAPI signal provides the nuclear contour. The three dimensional cell image was reconstructed by combining the contours of subsequent z-slices. (b) Statistics of the volumes of the cell (V), nucleus (V_N) and cytoplasm (V_C) for the three different cell lines (color coded). Notation: μ=mean, σ=standard deviation, cv= coefficient of variation, Δ^σ= the fraction of samples between $\mu - \sigma$ and $\mu + \sigma$, ρ= correlation between V_C and V_N, and $^*p < 0.001$ ($H_0 : \rho = 0$). (c-e) Scatterplots of V_C and V_N for the three different cell lines. Marginal histograms show the distribution of V_C (top) and V_N (right). The measured number of cells is given by n.

lines, indicating the dependence of expression levels on the gene location (see also the legend of fig. 3.5). Scaling of the mean and standard deviation are independent, indicating the independent regulation of the mean and variance of the expression level. We observe higher mRNA concentration in the cytoplasm than in the nucleus. Similar conclusions are drawn from the copy number data.

The comparison of the mRNA concentration and mRNA number data in terms of the coefficient of variation indicates that the concentration displays a smaller cell-to-cell variability across all cell lines. Thus, mRNA number noise overestimates the functional mRNA noise of the cells. Part of the reason why the concentration variability is smaller is the positive correlation of the cell volume and mRNA numbers per cell (fig. 3.16-3.18). This is indicated by fig. 3.1C, which shows a linear dependency of the mean mRNA copy numbers with volume. The dependency of the mean mRNA copy numbers on volume is proportional, such that a doubling in cell volume is accompanied by a doubling in the mRNA copy numbers. Remarkably, the cellular mRNA concentration conditional on the cell volume is constant, indicating homeostasis of the mRNA concentration (fig. 3.3C).

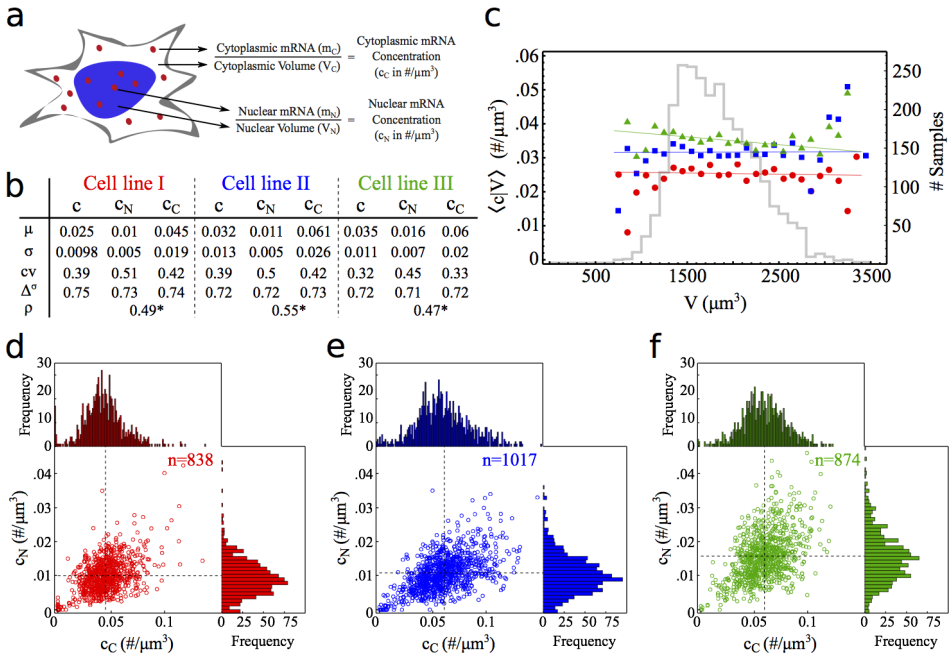

Figure 3.3: **Statistics of single-cell mRNA concentration data** (a) The previously obtained mRNA copy number (fig. 3.1) and volume data (fig. 3.2) were used to determine the concentration of mRNA in single cells (c), in the nucleus of these cells (c_N) and in the cytoplasm (c_C) (b) Statistics of the different mRNA concentrations for the three different cell lines (color coded). Notation: μ=mean, σ=standard deviation, cv= coefficient of variation, Δ^σ= the fraction of samples between $\mu - \sigma$ and $\mu + \sigma$, ρ= correlation between c_C and c_N, and $^*p < 0.001$ ($H_0 : \rho = 0$). (c) For a specific cell volume (V), the mean mRNA concentration is calculated. This conditional mean ($\langle c|V \rangle$) is constant with respect to volume. The gray histogram in the background shows the number of cells considered per volume bin (bin size = $100\mu m^3$). Higher counts indicate higher reliability of the corresponding determination of ($\langle c|V \rangle$). A least-squares linear fit is shown for all three cell lines, indicating mRNA concentration homeostasis. (d-f) Scatterplots of c_C and c_N for the three different cell lines. Marginal histograms show the distribution of c_C (top) and c_N (right). The given concentration ($\#/\mu^3$) can be converted to picomolars (pM) by multiplying with a conversion factor of 1660. The sample size is given by n. The bin size for the marginal histograms is 0.001 $\#/\mu m^3$.

3.2.4 The volume scaling of the mRNA concentration statistics explains the concentration variability

To address the differences between the concentration and the copy number noise we apply the law of total variance (section 3.3.2, fig. 3.8) to decompose the mRNA number and mRNA concentration noise each in a term that captures the volume-

	$\frac{var(m)}{\langle m\rangle^2}$	$\frac{\langle var(m\|V)\rangle}{\langle m\rangle^2}$	$\frac{var(\langle m\|V\rangle)}{\langle m\rangle^2}$	$\frac{var(c)}{\langle c\rangle^2}$	$\frac{\langle var(c\|V)\rangle}{\langle c\rangle^2}$	$\frac{var(\langle c\|V\rangle)}{\langle c\rangle^2}$	$\frac{var(V)}{\langle V\rangle^2}$
Cell line I	0.204	0.144	0.060	0.149	0.144	5.8×10^{-3}	0.064
Cell line II	0.208	0.149	0.059	0.156	0.152	3.3×10^{-3}	0.054
Cell line III	0.148	0.104	0.044	0.102	0.099	2.5×10^{-3}	0.053

Table 3.1: **The dependency of mRNA concentration and mRNA copy number noise on volume.** Normalized variances calculated from the experimental mRNA copy number, volume, and concentration statistics. Notation: $\langle x\rangle$ is the mean of the random variable x, $\langle x|y\rangle$ indicates the mean of x conditional on y, $var(x)$ denotes the variance of x and $var(x|y)$ denotes the conditional variance of x on y

dependence and another that quantifies gene-expression noise,

$$\frac{var(m)}{\langle m\rangle^2} = \overbrace{\frac{var(\langle m|V\rangle)}{\langle m\rangle^2}}^{\text{Volume-induced noise}} + \overbrace{\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}}^{\text{Gene-expression noise}}$$

$$\frac{var(c)}{\langle c\rangle^2} = \frac{var(\langle c|V\rangle)}{\langle c\rangle^2} + \frac{\langle var(c|V)\rangle}{\langle c\rangle^2}$$

Table 3.1 makes the same decomposition for the experimental data. It shows that the gene-expression noise term accounts for about 70% of the copy number noise and for over 95% of the concentration noise. Next we use theory to explain these observed normalized variances.

The volume-dependent noise contribution for the mRNA copy numbers can be estimated from theory using the experimentally-observed homeostasis relation: The homeostasis of the mRNA concentration resulted from the linear scaling, $\langle m|V\rangle = \alpha V$, with α as a positive constant and, as a consequence, we obtain $\langle m\rangle = \alpha\langle V\rangle$, $var(\langle m|V\rangle) = \alpha^2 var(V)$ and $var(\langle m|V\rangle)/\langle m\rangle^2 = var(V)/\langle V\rangle^2$. For an idealized model of cell growth, where cells divide at fixed intervals and into exactly equal halves, $var(V)/\langle V\rangle^2$ can be calculated to be approximately 0.04 (section 3.3.2.3), which provides a lower bound for $var(\langle m|V\rangle)/\langle m\rangle^2$. The overview of the variances of the experimental data in Table 3.1 shows that $var(\langle m|V\rangle)/\langle m\rangle^2$ is indeed very close to the theoretical estimate of 0.04. In the last column this value is directly calculated from the experimental volume distribution showing a slightly higher value. This is likely due to volume variation at fixed cell ages. The volume-dependent term explains 29.5%, 28.4%, and 30% of the copy number noise in cell line I, II, and III, respectively.

Due to mRNA concentration homeostasis, the volume-dependent term of the concentration noise equals zero; i.e. $var(\langle c|V\rangle)/\langle c\rangle^2 = 0$, which is also indicated by the experimental data (Column 6 of Table 3.1 and fig. 3.3C). Thus, the mRNA concentration noise is entirely determined by gene-expression noise.

The experimental data indicate that the gene-expression contributions to mRNA concentration and mRNA number noise are quite similar in absolute values (Column 2, 4 and 5 of Table 3.1). Since volume-derived noise in mRNA concentration is close to zero due to homeostasis, the mRNA concentration noise is approximately equal

to the gene-expression derived noise in mRNA copy numbers. Next, we will show that this is a direct consequence of the observed mRNA concentration homeostasis and, in addition, the scaling of the conditional mRNA number variance, $var(m|V)$, with cell volume.

At mRNA concentration homeostasis, the exact relationship between the gene-expression induced mRNA copy number and concentration noise is given by:

$$\frac{\langle var(c|V)\rangle}{\langle c\rangle^2} = \langle V\rangle^2\Big\langle\frac{1}{V^2}\Big\rangle\frac{\langle var(m|V)\rangle}{\langle m\rangle^2} + \frac{\sigma^2(1/V^2, var(m|V))}{\langle m\rangle^2/\langle V\rangle^2} \qquad \left(\approx \frac{var(c)}{\langle c\rangle^2}\right)$$

with $\sigma^2(x,y)$ denoting the covariance between x and y. Since we observe that $var(m|V)$ increases with volume (fig. 3.11), the covariance will be negative. Therefore, $\langle V\rangle^2\langle\frac{1}{V^2}\rangle$ is an upper bound for the relative deviation between the conditional noise in concentration and copy numbers in case of concentration homeostasis. This upper bound is reached when the covariance equals zero. From the volume probability distribution, $\langle V\rangle^2\langle\frac{1}{V^2}\rangle$ is estimated to be 1.12 (section 3.3.2.6). The deviation between $\frac{\langle var(c|V)\rangle}{\langle c\rangle^2}$ and $\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}$ requires the calculation of the covariance from the volume scaling relation of the mRNA copy number variance conditional on volume, i.e. from $\langle var(m|V)\rangle = \beta V^\gamma$. The experimental data (fig. 3.11) indicates that this scaling is maximally quadratic with volume. Theory predicts that in case of a linear dependence $\frac{\langle var(c|V)\rangle}{\langle c\rangle^2}$ is 4% higher than $\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}$ and in case of a quadratic dependence it is 4% lower. (Only when $\langle var(m|V)\rangle = \beta V^0$ the maximal deviation of 12% is achieved (section 3.3.2.6).) As can be seen from Table 3.1 (column 2 and 5) the relative difference between the two conditional noise terms indeed is close to the ±4% region as predicted by theory and the experimentally-observed volume scaling.

As a result of these relations, we can conclude that for our data the difference between mRNA concentration noise $\frac{var(c)}{\langle c\rangle^2}$ and mRNA copy number noise $\frac{var(m)}{\langle m\rangle^2}$ is dominated by the contribution of the volume dependent noise term which is zero for concentrations but equals values between 0.04 and 0.06 for copy numbers. Under conditions of mRNA concentration homeostasis this term is expected to be independent of the average expression level. The relative difference between mRNA copy number and concentration noise then depends on the magnitude of the volume independent noise contribution. For the three cell lines we investigated, this amounted to a 36%, 33%, and 45% difference between mRNA concentration and copy number noise. Thus, functional mRNA noise differs greatly from mRNA copy number noise indicating the importance of the combined measurement of molecule copy numbers and volumes of single cells.

3.2.5 Discussion

We studied three cell lines that only differed in the location of the same reporter construct controlled by a constitutive PGK-promoter. The differences in the mRNA statistics of these cell lines indicates gene-location dependency, which presumably results from the different chromatin states at the integration site. The mRNA copy

number and volume statistics that we measured resulted from pooled independent experiments. Individually these independent experiment lead to essentially identical distributions indicating that we sampled each time from a stationary growing cell population (fig. 3.5). Accordingly, the pooled volume data could be described by a volume distribution deriving from a stationary cell-age distribution for cells growing asynchronously at steady state (section 3.3.2.3, fig. 3.9).

We found that the mean mRNA copy number conditional on volume, $\langle m|V\rangle$, scaled linearly with volume, i.e. $\langle m|V\rangle = \alpha V$, which indicates homeostasis of the mRNA concentration as function of cell volume. This we interpret as a constant mRNA concentration while the cell volume grows. In addition, we found that the mRNA copy number variance conditional on the volume, $var(m|V)$, displayed a stronger than linear scaling with volume, i.e. $var(m|V) = \beta V^{\gamma}$ with $1 \leq \gamma \leq 2$. The latter scaling explained the difference between the volume dependent concentration and copy number noise, i.e. between $\frac{\langle var(c|V)\rangle}{\langle c\rangle^2}$ and $\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}$, which maximally amounts to a relative deviation of $\pm 4\%$ according to theory and in agreement with the experimental data. Taken together, these findings allow for a simple estimation of concentration noise based on copy number noise under conditions of mRNA concentration homeostasis: $\langle \delta^2 c\rangle/\langle c\rangle^2 \approx \langle \delta^2 m\rangle/\langle m\rangle^2 - 0.04$.

Our results indicate that constitutive gene expression is not completely understood at the level of a single cell. For homeostasis to occur during volume growth of the cell requires that $\langle m|V\rangle = \alpha V = \frac{k_s}{k_d} V$; with k_s and k_d as a zero-order and first-order rate constant for mRNA synthesis and degradation, respectively. (The life time of the mRNA is about 8 hours [Rowe 2007], i.e. much shorter than the generation time of about 24 hours.) In other words, either the transcription rate or the degradation rate of mRNA are volume dependent (or both such that the net effect leads to the proportionality of $\langle m|V\rangle$ with volume). This suggests a spontaneous coupling between the net rate of increase in the transcript numbers and the cell volume. It is not clear how this results from the combined influences of mRNA decay, replication, and cell volume dynamics. From this we conclude that constitutive gene expression is not completely understood at the level of a single cell.

Our data suggests a non-classical interpretation of constitutive gene expression, one where the synthesis (or degradation) rate scales with volume rather than a volume-independent constant transcription rate. In addition, the close to second-order scaling of $var(m|V)$ with volume hints at another not yet understood detail of constitutive gene expression. We observed these volume scalings with the same construct expressed from different genomic location at the level of single cells. The unexpected volume scaling is not an artifact of our reporter construct as underscored by several studies [Cookson 2010, Rosenfeld 2005, Sigal 2006, Cohen 2009]). Bennet et al. [Cookson 2010] found in yeast a peaked dependency of a green fluorescent protein expressed from a constitutive promoter as function of cell volume, which is indicative of constant synthesis and an accelerating growth of cell volume as function of the cell cycle. Similar data was reported for several human proteins [Sigal 2006]. Real-time monitoring of mRNA copy numbers (for instance by using MS2-labeling

[Fusco 2003]) and volume growth of single cells for a set of (classical) constitutive promoters should provide more information about the origins of the volume scaling relations of the mean and variance of the transcript copy numbers. A downside is that such studies would require the tracking of several hundreds of cell divisions to obtain robust statistics on the volume dependencies of copy-number statistics. To attain these robust statistics in our experiments, we used a confocal microscopy setup and studied almost a 1000 cells per cell line.

For inducible or cell-cycle dependent promoters, we expect much larger transient discrepancies between volume growth and mRNA synthesis. This would introduce much larger differences between the mRNA copy number and mRNA concentration noise than reported in this study, which is limited to a constitutive promoter. For inducible promoters we therefore expect that the assessment of mRNA concentration noise is even more relevant.

3.3 Supplemental Information

3.3.1 Materials and Methods

3.3.1.1 Cell lines and cell culture

Experiments were performed on human embryonic kidney cells (HEK293) with a single integration of a phosphoglycerate kinase (PGK) driven GFP gene construct, obtained from Gierman et. al. [Gierman 2007]. We analyzed three different cell lines with the integration at different genomic locations: Cell line I (HG19:chr1:225684028,within the ENAH gene), Cell line II (HG19:chr1:150379508, within the RPRD2 gene) and Cell line III (HG19:chr1:150664232, within the GOLPH3L gene). The cells were cultured in DMEM (Gibco®, 31965023) supplemented with 10% (v/v) fetal calf serum(Gibco®, 16140) and 100 units/ml Penicillin-Streptomycin (Gibco®, 15140). Incubation was at 37° C in a humidified 5% CO_2 atmosphere. Before any analysis the cells were grown for at least 2 weeks after thawing to achieve steady state cell growth and steady state expression statistics of the integrated construct.

3.3.1.2 Single Molecule RNA FISH

Samples were treated according to the Protocol for Adherent Mammalian Cell Lines for the Custom Stellaris™ FISH probes. Cells were cultured for 3 hours in Lab-Tek™ Chambered Coverglasses (Lab-Tek 155380) before fixation. EtOH permeabilization was done overnight at 4° C. For hybridization we used 125 nM probe in the hybridization buffer and incubated overnight at 37° C. Imaging was done without using anti-fade. The cells where counterstained with 5 ng/mL DAPI. The sequence of the probe targeting the eGFP insertion can be found in Supplemental Information section 3.3.7. The DNA probes were coupled to CAL Fluor© Red 590 fluorophores by the manufacturer (Biosearch Technologies, Inc.).

3.3.1.3 Image acquisition

Samples were imaged using a Nikon Ti-E scanning laser confocal inverted microscope (A1) with 60x oil objective in tandem with Nikon NIS-Elements imaging software. Excitation was by 561.5 nm diode-pumped solid state and 402.1 nm diode lasers. Detection was via 595-50 nm and 450-50 nm filters, respectively. Optical sections were captured at 0.300 μm intervals and a resolution of 256 by 256 pixels and zoom factor of 6.8, resulting in a voxel size of 0.0047 μm^3 (0.1243 μm by 0.1243 μm by 0.3 μm). Four times averaging was used to reduce photon and camera noise.

3.3.1.4 mRNA counts

Image analysis software was adapted from Raj et. al. [Raj 2008a]. Images are filtered with a semi three-dimensional Laplacian of Gaussian (LoG) filter which removes noise and enhances the signal to noise ratio (filter width=1.5). The number

of mRNA spots was found by applying a threshold for which the number of mRNA was least sensitive to changes in this threshold. The threshold was determined by using a window function calculating the average spot numbers over 7 constitutive thresholds divided by the sum of the standard deviation of these spot counts and a constant α (=10) [Itzkovitz 2012a]. For each cell line multiple experiments were performed to obtain good statistics of mRNA copy numbers and volumes data. Each single experiment spanned several hours. Figures 3.4 and 3.5 show that there is no indication for a decrease in sample quality during the course of imaging and that the data collected on different days are nearly identical.

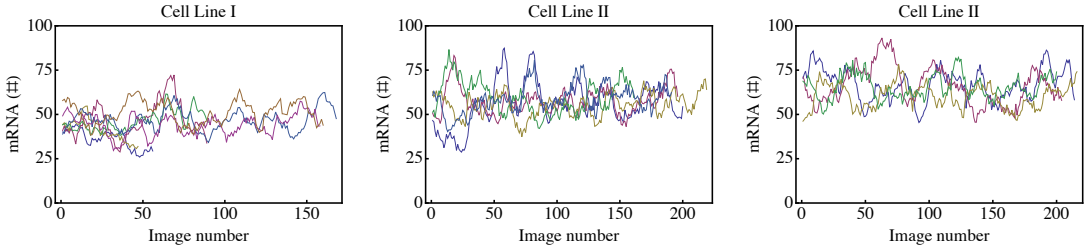

Figure 3.4: **Confirmation of the absence of scan time bias in the mRNA count.** During image acquisition the quality of the sample might decrease. To test whether this is the case in our data set we calculated the moving averages of the number of identified mRNAs. Here we show the moving averages for all experiments performed on the three different cell lines (from left to right; Cell line I, Cell line II and Cell line III). If the colored lines, which show the data of a single experiment, have a constant tendency of decrease or increase the sample quality is not stable during image acquisition. Although this might be the case in some experiments (Cell line I, smallest samples), overall we observe no influence of the acquisition time on the identified number of mRNA spots in the images.

3.3.1.5 Volume measurements

To obtain the volume of the cells we analyzed both recorded channels individually, the DAPI and the probe, using MATLAB Release 2012b. The following operations are performed on all individual z-slices of the images:

1. Median filtering (20 by 20 pixels).
2. Image thresholding (graythresh, Otsu's method [Otsu 1979]).
3. Fill image regions and holes.
4. Morphological closing with a disk (radius = 4 pixels).

After these operations the three dimensional image is reconstructed. From processed images of the DAPI channel the nuclear volume can be obtained. The cell volume is defined by the presence of signal from either the nucleus, the cytoplasm or both. The cytoplasmic size is given by the difference between nuclear and cell volume.

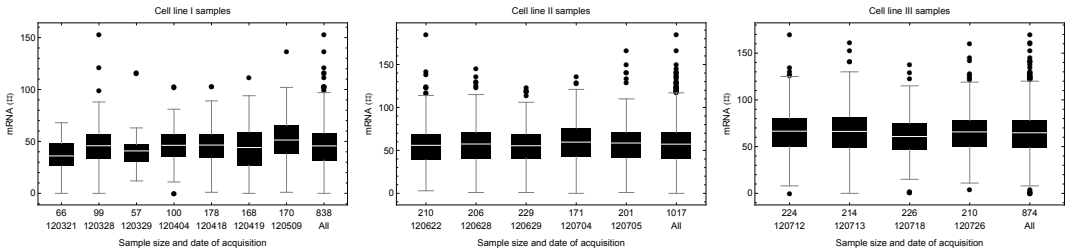

Figure 3.5: **Box-whisker plots of the individual experiments indicate almost no day to day variation between single experiments.** To accurately measure the mRNA probability distributions we need multiple experiments for each cell line. The three box-whisker plots show the mRNA counts of the single experiments. There is no indication that the individual samples of Cell line II ($p = 0.52$) and Cell line III ($p = 0.16$) are not drawn from the same distribution. There is higher variation present between the individual samples of Cell Line I (p=3 10^{-7}), p-values are obtained by Kruskal-Wallis tests, which assumes that the data has a symmetric distribution. This variation might be due to the longer cell culturing period during sample acquisition in combination with the small sample sizes of the Cell line I experiments. The individual experiments are annotated with the number of samples per experiment at the x-axes and date at which the experiment is preformed (yymmdd). The last box shows all single experiments of the specific cell line combined. To test whether the different cell lines had different gene expression characteristics we performed an ANOVA and compared the three cell lines based on the statistics of the individual experiments. The mean ($p < 0.0001$), the coefficient of variation ($p < 0.031$), the noise ($p < 0.05$), and the Fano factor ($p < 0.04$) all appear to differ for the different cell lines. For the mRNA concentration data similar results are obtained; the mean ($p < 0.0001$), the coefficient of variation ($p < 0.013$), the noise ($p < 0.016$).

The obtained pixel size was multiplied by the voxel size of $0.0047\mu m^3$ to provide the cell size in μm^3. The resulting distributions are shown in figure 3.6.

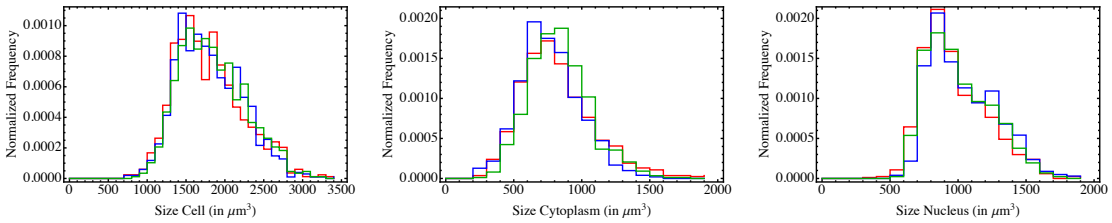

Figure 3.6: **Volume measurements provide a similar volume distribution for the three different cell lines.** Histograms showing the volume distribution of the cells, of the cytoplasm and of the nucleus. The three different cell lines are indicated by the different colors (red=cell line I, blue=cell line II, and green=cell line III). The bin-size is 100 μm^3. The volume distributions of the different cell lines are expected to be the same since the cells differ only in the integration site of the GFP-construct. We tested wether the sample means of the different cell lines originate from the same distribution with ANOVA. ($V; p = 0.65$, $V_N; p = 0.12$, $V_C; p = 0.24$)

3.3.1.6 DAPI and nuclear lamina staining

The nuclear counterstain 4'-6-Diamidino-2-phenylindole (DAPI) is used to identify the localization of the mRNA's as either cytoplasmic or nuclear. To check how well the DAPI staining confines the nucleus, we compare DAPI staining with a mCherry tagged nuclear lamina protein (LMNB). Cells were transiently transfected with Lipofectamin 2000, the ratio of lipofectamin to DNA was 2.5, according to the manufacturer instructions. Images were acquired using 402.1 nm and 591.5 nm lasers for excitation and detection was via a 450-50 nm band pass and 605 nm long pass filters, respectively.

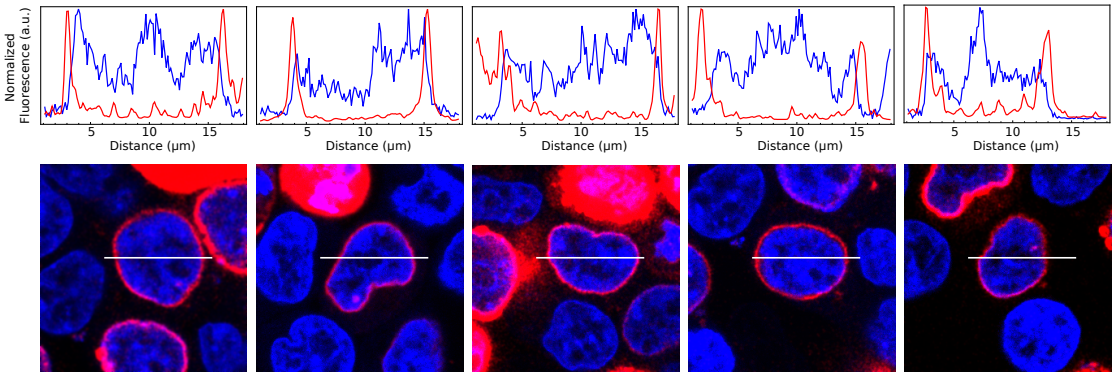

Figure 3.7: **Lamina staining confirms that DAPI reliably distinguishes the nuclear boundary.** Overall the DAPI staining (Blue signal) coincides with the lamin staining (Red signal), as can be seen by the profile above the microscopy images. Images are one single z-slice. The profiles show that where the lamin staining peaks (red line) the dapi staining (blue line) increases or decreases rapidly.

3.3.2 The law of total variance for the copy numbers and concentrations of mRNA

3.3.2.1 Law of total variance explained

The law of total variance can be derived from the definition of the variance, $var(m) = \langle m^2 \rangle - \langle m \rangle^2$, and the law of total expectation, $\langle m \rangle = \langle \langle m|V \rangle \rangle$.

$$\begin{aligned} var(m) &= \langle m^2 \rangle - \langle m \rangle^2 \\ &= \langle \langle m^2|V \rangle \rangle - \langle \langle m|V \rangle \rangle^2 \\ &= \langle var(m|V) + \langle m|V \rangle^2 \rangle - \langle \langle m|V \rangle \rangle^2 \\ &= \langle var(m|V) \rangle + \langle \langle m|V \rangle^2 \rangle - \langle \langle m|V \rangle \rangle^2 \\ &= \langle var(m|V) \rangle + var(\langle m|V \rangle) \end{aligned} \tag{3.1}$$

We can do the same for the concentration of mRNA (c) and arrive at:

$$var(c) = \langle var(c|V) \rangle + var(\langle c|V \rangle) \tag{3.2}$$

In this equation, $\langle var(c|V) \rangle$ is the mean of the variance at fixed V. The second term, $var(\langle c|V \rangle)$, is the variance of the conditional means and represents the variance caused by the change in V (figure 3.8).

3.3.2.2 Deriving $\frac{var(\langle m|V \rangle)}{\langle m \rangle^2} = \frac{var(V)}{\langle V \rangle^2}$ in case of homeostasis

Homeostasis requires that the average mRNA copy number per cell at a given volume, i.e. $\langle m|V \rangle$ scales linearly with the volume, i.e. $\langle m|V \rangle = \alpha V$. As a consequence, we obtain that $\langle m \rangle = \langle \langle m|V \rangle \rangle_V = \alpha \langle V \rangle$ and $var(\langle m|V \rangle) = var(\alpha V) = \alpha^2 var(V)$. This means that in case of perfect homeostasis: $\frac{var(\langle m|V \rangle)}{\langle m \rangle^2} = \frac{var(V)}{\langle V \rangle^2}$. In the next section we will derive that $\frac{var(V)}{\langle V \rangle^2} \approx 0.04$.

3.3.2.3 Estimation of a lower bound for $\frac{var(\langle m|V \rangle)}{\langle m \rangle^2}$ assuming steady-state exponential growth of the cells

Here we estimate the noise in the volume distribution for an idealized model of cell division. We assume that cells divide at fixed intervals T and divide into exactly equal halves. Furthermore we assume that the volume is a deterministic function of cell age (denoted by a): $V = V(a) = V_0 e^{\mu a}$ with μ as the specific (exponential) growth rate. The daughter cell volume equals $V(0) = V_0$ and the mother volume at division equals $V(T) = 2V_0$. Hence, $0 \leq a \leq T$ with $T = \ln 2/\mu$. At balanced growth, the distribution of cell ages for this model is described by a so-called ideal age distribution [Sueoka 1965] equal to

$$u(a) = \mu 2^{1-a\mu/\ln(2)} \qquad a \in [0, \ln(2)/\mu] \tag{3.3}$$

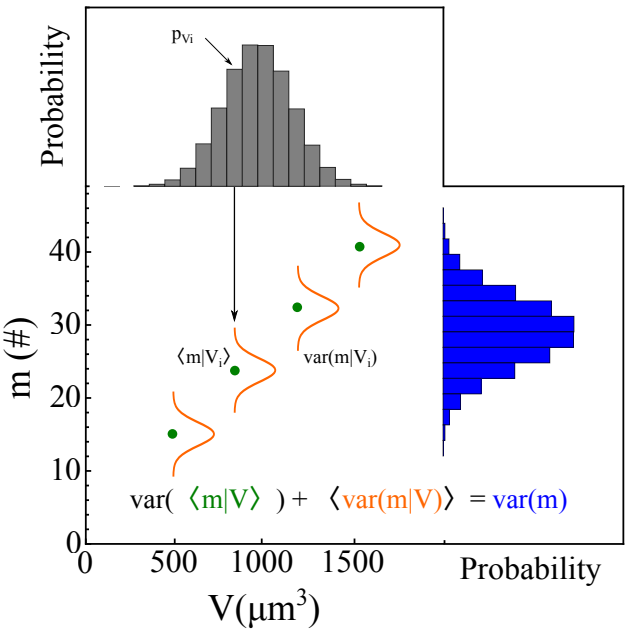

Figure 3.8: **Visualization of the law of total variance.** For all samples there are two measurements, in this case called volume (V) (x-axis) and the number of mRNA molecules (m) (y-axis). The marginal distributions show the probability distributions of both variables, V (top) and m (right). The samples within each group (V_i) have a distribution of m of their own ($m|V$), shown in orange, with their means indicated in green. The variance in m is the sum of the variance over the conditional means, $var\langle m|V\rangle$, and the mean over the conditional the variances, $\langle var(m|V)\rangle$.

The probability distribution of volumes, $g(V)$, can be derived directly from the age distribution using the change of variable technique:

$$a = \ln(V/V_0)/\mu \tag{3.4}$$

$$g(V) = \left|\frac{\partial}{\partial V}(\ln(V/V_0)/\mu)\right| u(\ln(V/V_0)/\mu) \tag{3.5}$$

$$= 2V_0/V^2 \qquad V \in [V_0, 2V_0] \tag{3.6}$$

The noise in this distribution is given by:

$$\frac{var(V)}{\langle V\rangle^2} = \frac{2}{\ln(4)^2} - 1 \approx 0.04 \tag{3.7}$$

This distribution is shown in figure 3.9. The assumption of deterministic interdivision times leads to the unrealistic discontinuous definition of the distribution function. A better fit to the experimentally observed distributions is obtained when the volumes at which cells divide (and those of newborn cells) are allowed to have

some variation around the mean value. With scaled, symmetric beta distributions for the cell volume at cell birth and division, each with $CV = 0.1$ a much better fit to the experimental data is obtained (see figure 3.9). In this case, the distribution of volumes can be obtained from the equations deduced by Collins and Richmond [Collins 1962] and the corresponding cell age and inter division time distributions can be derived as well [Painter 1968], resulting in a coefficient of variation of 20% for the interdivision time distribution.

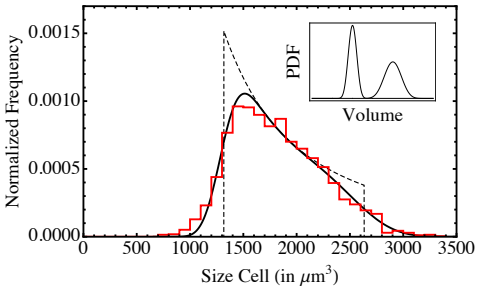

Figure 3.9: **Correspondence of the theoretical and experimental volume distribution** for either an ideal cell age distribution (dashed line) or for a population where volumes at birth and at division follow scaled, symmetric beta distributions with coefficients of variation of 10% (solid black line). The experimentally obtained volume measurements are shown in red. The inset shows the beta distributions used for volumes of dividing and newborn cells (distribution functions: $f_b(V) = \frac{\left(\frac{3}{2}-V\right)^{-1+a}\left(-\frac{1}{2}+V\right)^{-1+a}}{\text{Beta}[a,a]}, V \in [0.5V_0, 1.5V_0]$ and $f_m(V) = \frac{1}{2}\frac{\left(\frac{3}{2}-\frac{V}{2}\right)^{-1+a}\left(-\frac{1}{2}+\frac{V}{2}\right)^{-1+a}}{\text{Beta}[a,a]}, V \in [V_0, 3V_0]$, $a = 12$).

3.3.2.4 General relation between $\frac{\langle var(c|V)\rangle}{\langle c\rangle^2}$ and $\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}$

As explained in the main text we found from the experimental data that mRNA concentration and copy number noise differ. In this section we will derive a relation between $\frac{\langle var(c|V)\rangle}{\langle c\rangle^2}$ and $\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}$.

The concentration is defined as $c = \frac{m}{V}$, which leads to the following additional relations,

$$var(c|V) = var(\frac{m}{V}|V) = \frac{1}{V^2}var(m|V) \tag{3.8}$$

$$\langle var(c|V)\rangle = \left\langle \frac{1}{V^2}var(m|V)\right\rangle \tag{3.9}$$

$$= \left\langle \frac{1}{V^2}\right\rangle\langle var(m|V)\rangle + \underbrace{\sigma^2(1/V^2, var(m|V))}_{\substack{\text{covariance between the squared inverse} \\ \text{volume and the conditional copy} \\ \text{number variance}}} \tag{3.10}$$

$$\frac{\langle var(c|V)\rangle}{\langle c\rangle^2} = \frac{\langle \frac{1}{V^2}\rangle\langle var(m|V)\rangle + \sigma^2(1/V^2, var(m|V))}{\langle m/V\rangle^2} \tag{3.11}$$

The relation used to go from the second to the third equation is $\langle\alpha\beta\rangle = \langle\alpha\rangle\langle\beta\rangle + \sigma^2(\alpha,\beta)$ with $\sigma^2(\alpha,\beta)$ as the covariance between the random variables α and β.

3.3.2.5 Simplification of the relation between $\frac{\langle var(c|V)\rangle}{\langle c\rangle^2}$ and $\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}$ in case of homeostasis

In case of homeostasis we can simplify equation 3.11. Homeostasis means that $\langle m|V\rangle$ scales proportional with V and, as a consequence, that the concentration is independent of volume and fixed: $\langle c|V\rangle = \alpha$. Thus, $\langle m|V\rangle = \alpha V$ and $\langle c|V\rangle = \langle m|V\rangle/V$.

First we will show that under these homeostasis conditions $\langle m/V\rangle^2 = \langle m\rangle^2/\langle V\rangle^2$. The definition of homeostasis implies,

$$\langle c|V\rangle = \frac{\langle m|V\rangle}{V} = \alpha \tag{3.12}$$

Averaging this equation over volume leads to the relation $\langle m\rangle = \alpha\langle V\rangle$. Averaging $\langle c|V\rangle$ over the whole volume distribution gives:

$$\langle\langle c|V\rangle\rangle_V = \langle c\rangle = \left\langle\frac{m}{V}\right\rangle = \langle\alpha\rangle = \alpha \tag{3.13}$$

Hence, in case of homeostasis $\langle\frac{m}{V}\rangle = \frac{\langle m\rangle}{\langle V\rangle}$. Substituting the relation $\langle m/V\rangle^2 = \langle m\rangle^2/\langle V\rangle^2$ into equation 3.11 gives:

$$\frac{\langle var(c|V)\rangle}{\langle c\rangle^2} = \langle V\rangle^2\langle\frac{1}{V^2}\rangle\frac{\langle var(m|V)\rangle}{\langle m\rangle^2} + \frac{\sigma^2(1/V^2, var(m|V))}{\langle m\rangle^2/\langle V\rangle^2} \tag{3.14}$$

3.3.2.6 Simplifying the relation between $\frac{\langle var(c|V)\rangle}{\langle c\rangle^2}$ and $\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}$ from volume scaling relations

To estimate the magnitude of the covariance term in equation 3.14 we approximate $var(m|V)$ as a polynomial in V as this scaling is also observed in our data (figure 3.11).

$$var(m|V) = a_0 + a_1V + a_2V^2 \tag{3.15}$$

Averaging this equation over the volume gives,

$$\langle var(m|V)\rangle = \langle a_0 + a_1V + a_2V^2\rangle = a_0 + a_1\langle V\rangle + a_2\langle V^2\rangle \tag{3.16}$$

Using the volume distribution, $g(V)$ we can calculate the covariance between $var(m|V)$ and $1/V^2$:

$$\begin{aligned}
\sigma^2 & (\frac{1}{V^2}, var(m|V)) \\
&= \int_{V_0}^{2V_0} g(V)\left(\frac{1}{V^2} - \left\langle\frac{1}{V^2}\right\rangle\right)(a_0 + a_1V + a_2V^2 - \langle a_0 + a_1V + a_2V^2\rangle)dV \\
&= a_1\left(\left\langle\frac{1}{V}\right\rangle - \langle V\rangle\left\langle\frac{1}{V^2}\right\rangle\right) + a_2\left(1 - \langle V^2\rangle\left\langle\frac{1}{V^2}\right\rangle\right).
\end{aligned} \tag{3.17}$$

Combining equations 3.14, 3.16 and 3.17 yields:

$$\frac{\langle var(c|V)\rangle}{\langle c\rangle^2} = \underbrace{\left(\langle V\rangle^2\langle\frac{1}{V^2}\rangle\right)}_{\substack{\approx\ 1.12\\ \text{(eq. 3.19)}}}\frac{a_0}{\langle m\rangle^2} + \underbrace{\left(\langle V\rangle\langle\frac{1}{V}\rangle\right)}_{\substack{\approx\ 1.04\\ \text{(eq. 3.20)}}}\frac{a_1\langle V\rangle}{\langle m\rangle^2} + \underbrace{\frac{\langle V\rangle^2}{\langle V^2\rangle}}_{\substack{\approx\ 0.96\\ \text{(eq. 3.21)}}}\frac{a_2\langle V^2\rangle}{\langle m\rangle^2} \quad (3.18)$$

where the approximated values are calculated using the volume distribution (eq. 3.6):

$$\langle V\rangle^2\langle\frac{1}{V^2}\rangle = \left(\int_{V_0}^{2V_0} Vg(V)dV\right)^2 \int_{V_0}^{2V_0}\frac{1}{V^2}g(V)dV = \frac{7\ln(4)^2}{12} \approx 1.12 \quad (3.19)$$

$$\langle V\rangle\langle\frac{1}{V}\rangle = \int_{V_0}^{2V_0} Vg(V)dV \int_{V_0}^{2V_0}\frac{1}{V}g(V)dV = \frac{3\ln(4)}{4} \approx 1.04 \quad (3.20)$$

$$\frac{\langle V\rangle^2}{\langle V^2\rangle} = \frac{\left(\int_{V_0}^{2V_0} Vg(V)dV\right)^2}{\int_{V_0}^{2V_0} V^2g(V)dV} = \frac{(\ln(4))^2}{2} \approx 0.96 \quad (3.21)$$

With the experimental volume distribution (calculated based on the pooled volume data from all three cell-lines) these values become:

$$\langle V\rangle^2\langle\frac{1}{V^2}\rangle = 1.18 \quad (3.22)$$

$$\langle V\rangle\langle\frac{1}{V}\rangle = 1.06 \quad (3.23)$$

$$\frac{\langle V\rangle^2}{\langle V^2\rangle} = 0.95 \quad (3.24)$$

On basis of the theory we can distinguish three regimes:

1. No scaling of $var(m|V)$ with volume: $var(m|V) = a_0$ then $\frac{\langle var(c|V)\rangle}{\langle c\rangle^2} = 1.12\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}$; a 12% larger mRNA concentration noise than copy number noise.

2. Linear scaling of $var(m|V)$ with volume: $var(m|V) = a_1V$ then $\frac{\langle var(c|V)\rangle}{\langle c\rangle^2} = 1.04\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}$; a 4% larger mRNA concentration noise than copy number noise.

3. Quadratic scaling of $var(m|V)$ with volume: $var(m|V) = a_2V^2$ then $\frac{\langle var(c|V)\rangle}{\langle c\rangle^2} = 0.96\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}$; a 4% smaller mNRA concentration noise than copy number noise.

Our data do not distinguish between a linear and a quadratic dependence of conditional copy number variance on volume, both regimes give decent fits (figure 3.11). Hence, a discrepancy between $\frac{\langle var(c|V)\rangle}{\langle c\rangle^2}$ and $\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}$ of $\pm 4\%$ is expected on the basis of the theoretical analysis. This agrees quite well with the difference in the

experimental values of $\frac{\langle var(c|V)\rangle}{\langle c\rangle^2}$ and $\frac{\langle var(m|V)\rangle}{\langle m\rangle^2}$ (Table 1, main text); i.e. for cell line I, II and III we find respectively 0%, 2% and 4.8% difference, which agrees very well with the theoretical estimates for this discrepancy to lie between ±4%. However, the noise in mRNA copy number and concentration show a larger discrepancy, i.e. of 36% $((0.204 - 0.149)/0.149 \times 100\%)$, 33% $((0.208 - 0.156)/0.156 \times 100\%)$, and 45% $((0.148 - 0.102)/0.102 \times 100\%)$ for cell line I, II and III, respectively. These differences are due to the scaling of the copy number with cell volume due to the steady-state growth of the cells.

3.3.3 Average mRNA copy numbers correlate well with protein expression

The GFP-protein expression levels of the integrated construct are compared to the GFP-mRNA copy number expression. The protein expression levels of the cell lines originate from Gierman et. al. [Gierman 2007]. As expected higher mean mRNA copy numbers corresponds to higher protein levels (fig. 3.10).

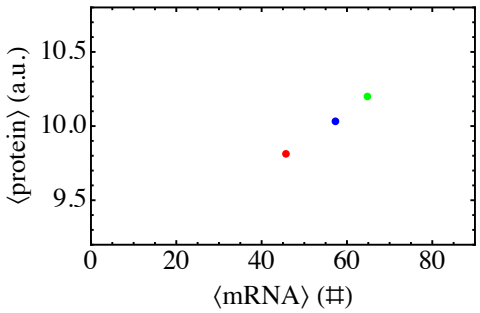

Figure 3.10: **Apparent correlation between the mRNA and protein expression levels.** Cell line I is shown in red, Cell line II in blue and Cell line III in green.

3.3.4 Concentration homeostasis and proportionality of the mRNA copy numbers as function of volume

The conditional variances of m and c are given in figure 3.11. The decomposition of averages and variances as function of volume for the nucleus and cytoplasm of the mRNA copy numbers (figure 3.12) show similar proportionality as observed for the whole cell data. Homeostasis of mRNA concentrations is also observed at the nuclear (figure 3.13A) and cytoplasmic level (figure 3.13C).

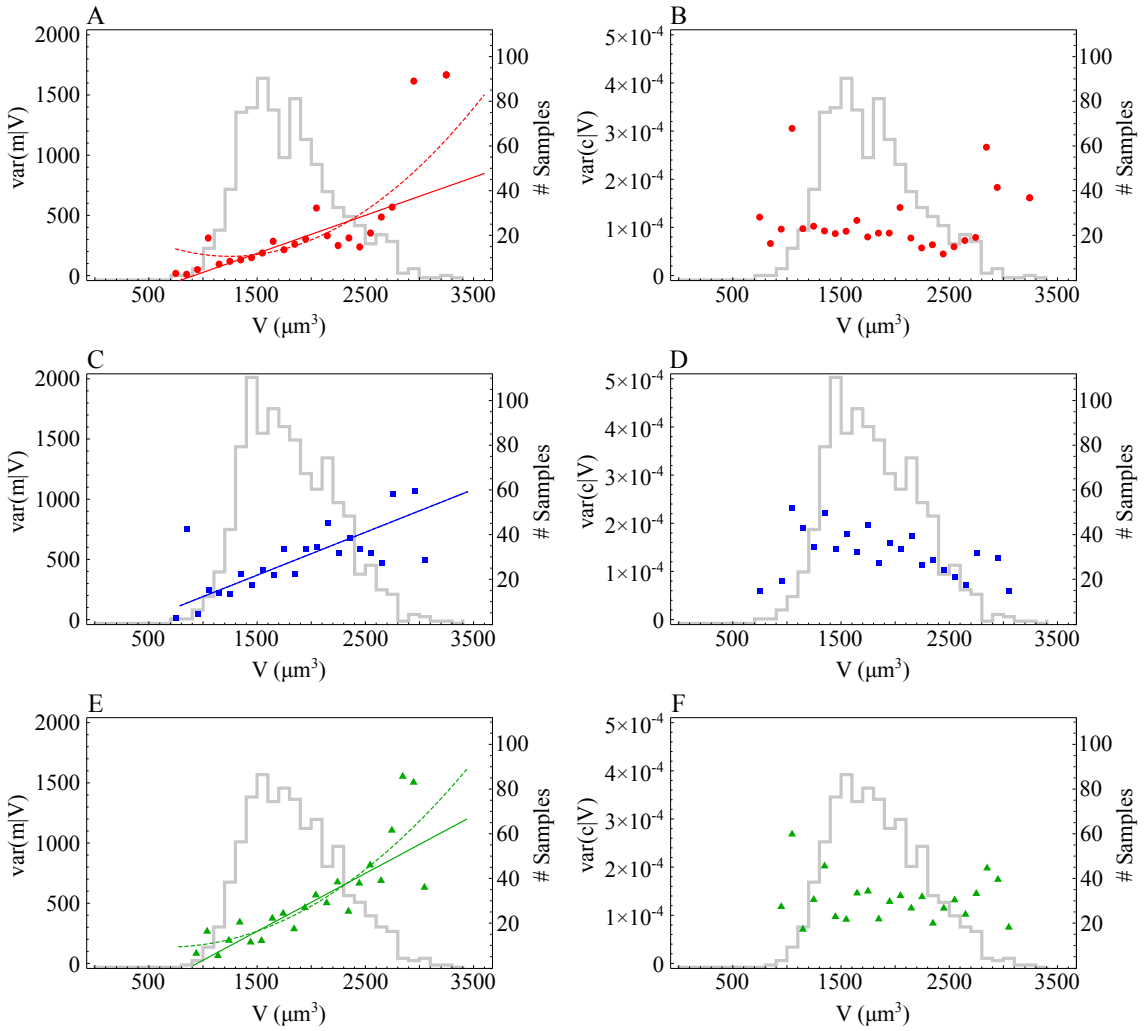

Figure 3.11: **Conditional variances of copy numbers and concentrations.** A, C and E give the variance in mRNA number conditioned on the cell volume ($var(m|V)$). B, D and F give the variance in the conditioned concentration ($var(c|V)$). The three different cell lines are annotated by the different colors: Cell line I is shown in red, Cell line II in blue and Cell line III in green. The gray plots in the background shows the number of data points per volume bin. The lower the count in one bin, the lower the reliability of the measurement of the data within the bin. The copy number data is fitted with a linear (solid, $a + bV$) and polynomial (dashed, $a + bV + cV^2$ with $c > 0$) fit averaging for the number of data points in each volume bin. For Cell line II (C) the linear and polynomial fit are overlapping.

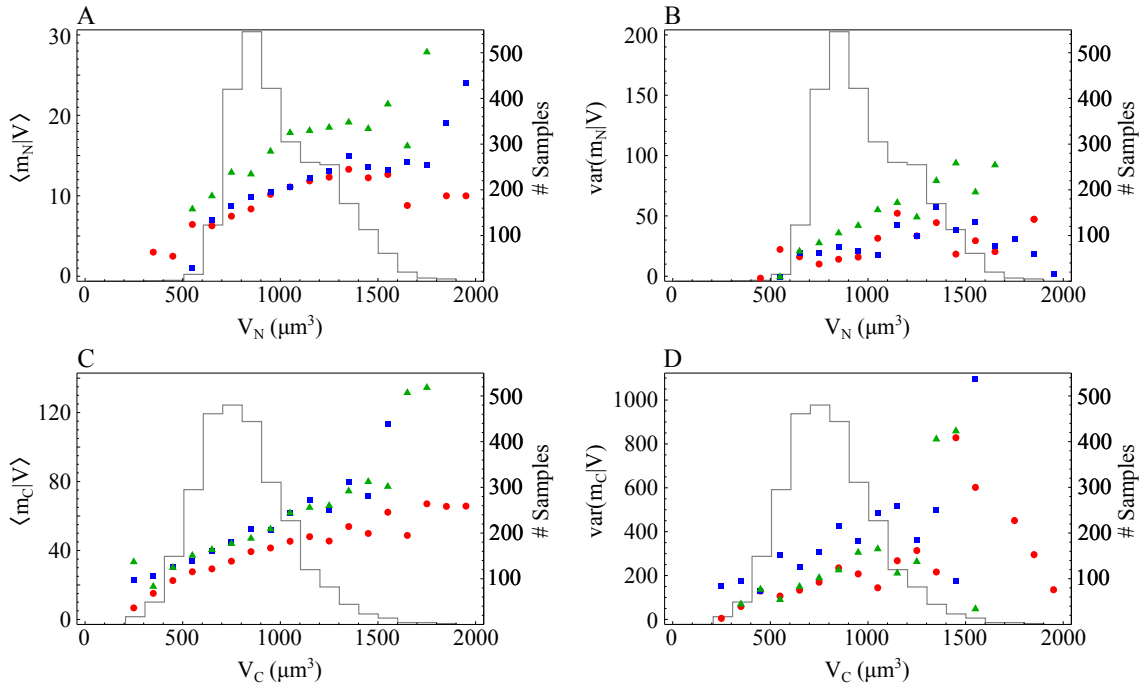

Figure 3.12: **Conditional averages and variances in the mRNA copy number in the nucleus and cytoplasm.** The three different cell lines are annotated by the different colors: Cell line I is shown in red, Cell line II in blue and Cell line III in green. The gray plots in the background shows the number of data points per volume bin for either the nucleus (A and B) or the cytoplasm (C and D).

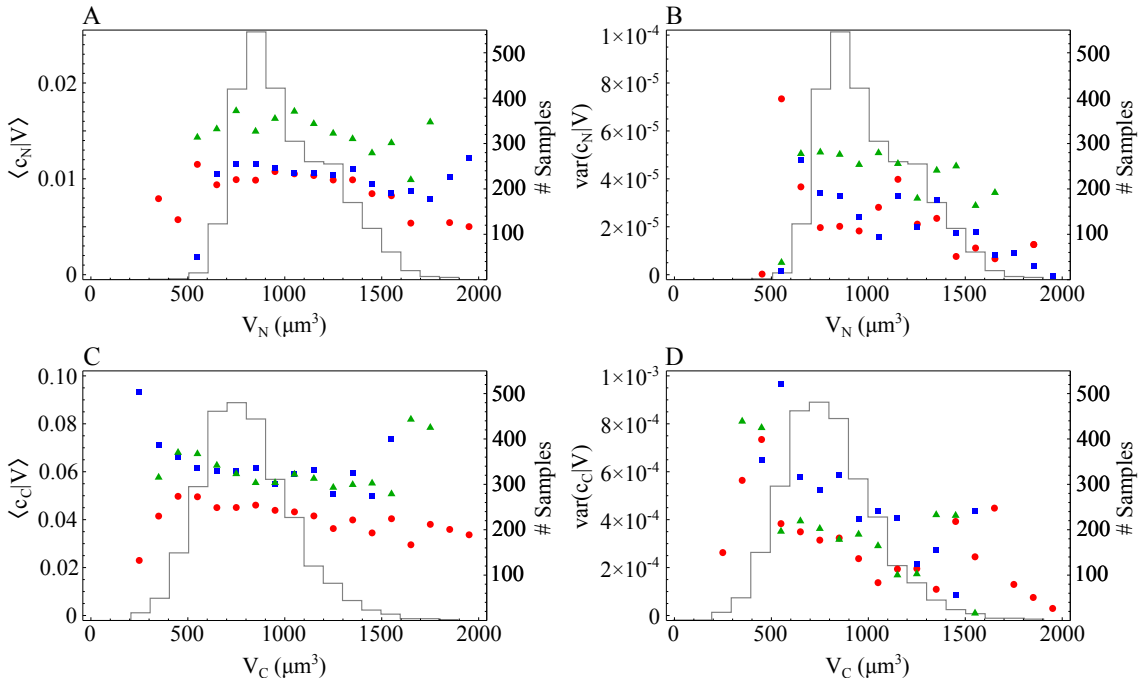

Figure 3.13: **Conditional averages and variances in the mRNA concentration in the nucleus and cytoplasm.** The three different cell lines are annotated by the different colors: Cell line I is shown in red, Cell line II in blue and Cell line III in green. The gray plots in the background shows the number of data points per volume bin for either the nucleus (A end B) or the cytoplasm (C and D).

3.3.5 Summary of the distribution statistics

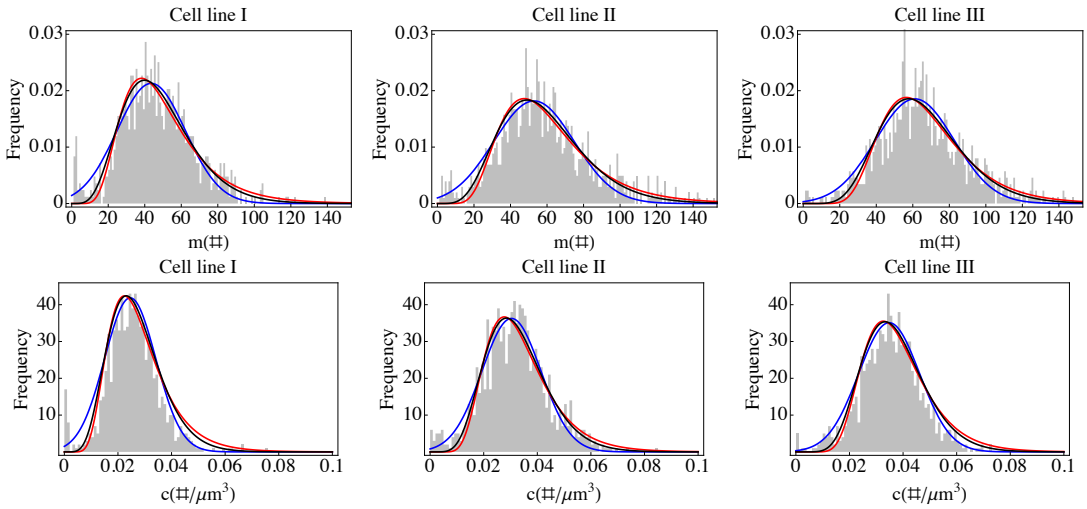

Figure 3.14: **The whole - cell mRNA expression distributions in terms of mRNA copy number (upper row) and mRNA concentration (second row) of the three different cell lines (the different columns)**.The measured expression distribution is shown in gray. Fits of the data with a normal distribution (blue lines), lognormal distribution (red lines) and gamma distribution (black lines) are shown.

A	Cell line I			Cell line II			Cell line III		
	m	m_N	m_C	m	m_N	m_C	m	m_N	m_C
μ	45.7	9.7	36	57.3	11.2	46.1	64.8	15.7	49.1
σ	20.7	5.40	17.1	26.1	5.91	21.8	25	7.67	19.4
σ^2	427	29.2	292.7	681.7	35	474.9	620.6	58.9	376.9
cv	.452	.556	.476	.456	.529	.473	.384	.488	.396
fano	9.35	3	8.13	11.9	3.13	10.3	9.58	3.75	7.68
noise	.205	.309	.226	.208	.28	.223	.148	.238	.156
B	V	V_N	V_C	V	V_N	V_C	V	V_N	V_C
μ	1800	979	822	1808	1040	768	1843	1008	835
σ	456	238	274	419	250	221	423	235	223
σ^2	$21\ 10^4$	$56\ 10^3$	$75\ 10^3$	$18\ 10^4$	$62\ 10^3$	$49\ 10^3$	$18\ 10^4$	$55\ 10^3$	$50\ 10^3$
cv	.253	.243	.334	.231	.24	.287	.229	.232	.267
noise	.064	.059	.112	.054	.058	.083	.053	.054	.071
C	c	c_N	c_C	c	c_N	c_C	c	c_N	c_C
μ	.0254	.01	.0447	.0317	.0109	.0609	.0354	.0158	.0596
σ	.0098	.0051	.0186	.0125	.0054	.0255	.0113	.0071	.0197
σ^2	$1\ 10^{-4}$	$3\ 10^{-5}$	$4\ 10^{-4}$	$1.6\ 10^{-4}$	$3\ 10^{-5}$	$7\ 10^{-4}$	$1.3\ 10^{-4}$	$5\ 10^{-5}$	$4\ 10^{-4}$
cv	.387	.509	.416	.395	.499	.419	.319	.449	.331
noise	.15	.259	.173	.156	.25	.176	.102	.202	.109

Figure 3.15: **Complete data table of the measured distribution statistics.** The mean (μ),standard deviation (σ), variance (σ^2), coefficient of variation (cv) and noise are shown. For the mRNA copy number measurements, the fano factor is given.

3.3.6 Correlations All vs All

For all measurements in the data set of each cell line we calculated the correlation coefficient with all other measurements. The probabilities of the correlation are determined by calculating the t statistic by $t = r\sqrt{\frac{n-2}{1-r^2}}$ where r is the sample correlation, n the sample size. Corresponding p-values are are derived from a t distribution with $n-2$ degrees of freedom. This test returns the probability of the observed correlation under the null hypothesis $H_0 : r = 0$. Figures 3.16, 3.17, 3.18 show that the observed correlations are all significantly different from zero, exceptions are the correlations between the volumes and the concentration measurements. These correlations are found to be around zero and with much higher p-values. The p-values reported are not corrected for multiple testing. The measurements of the total cell characteristics are not independent measurements when compared to the same measurement of the cytoplasm or nucleus.

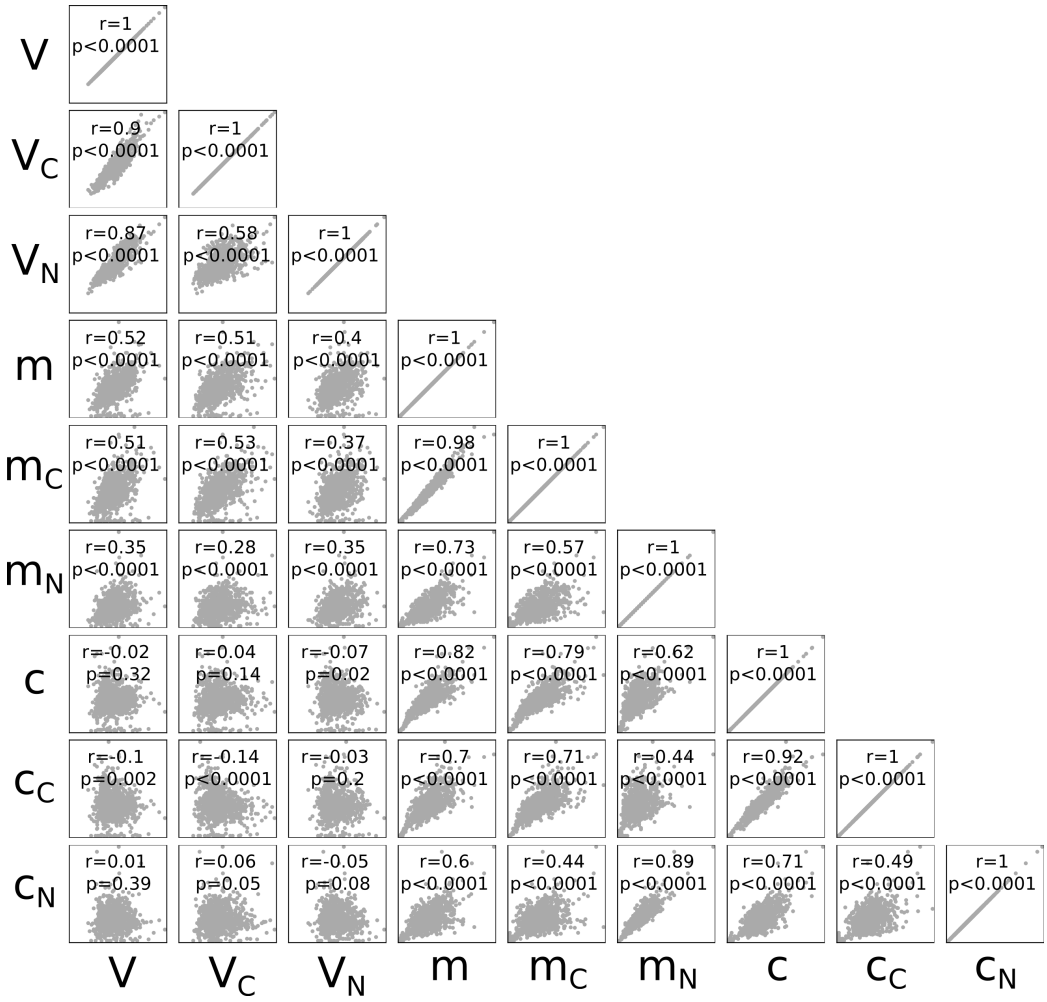

Figure 3.16: **Correlations within the data of Cell line I.** Within each plot the correlation coefficient(r) and the corresponding p-value are shown. The boxed region shows that the concentrations have low correlations with the volume measurements.

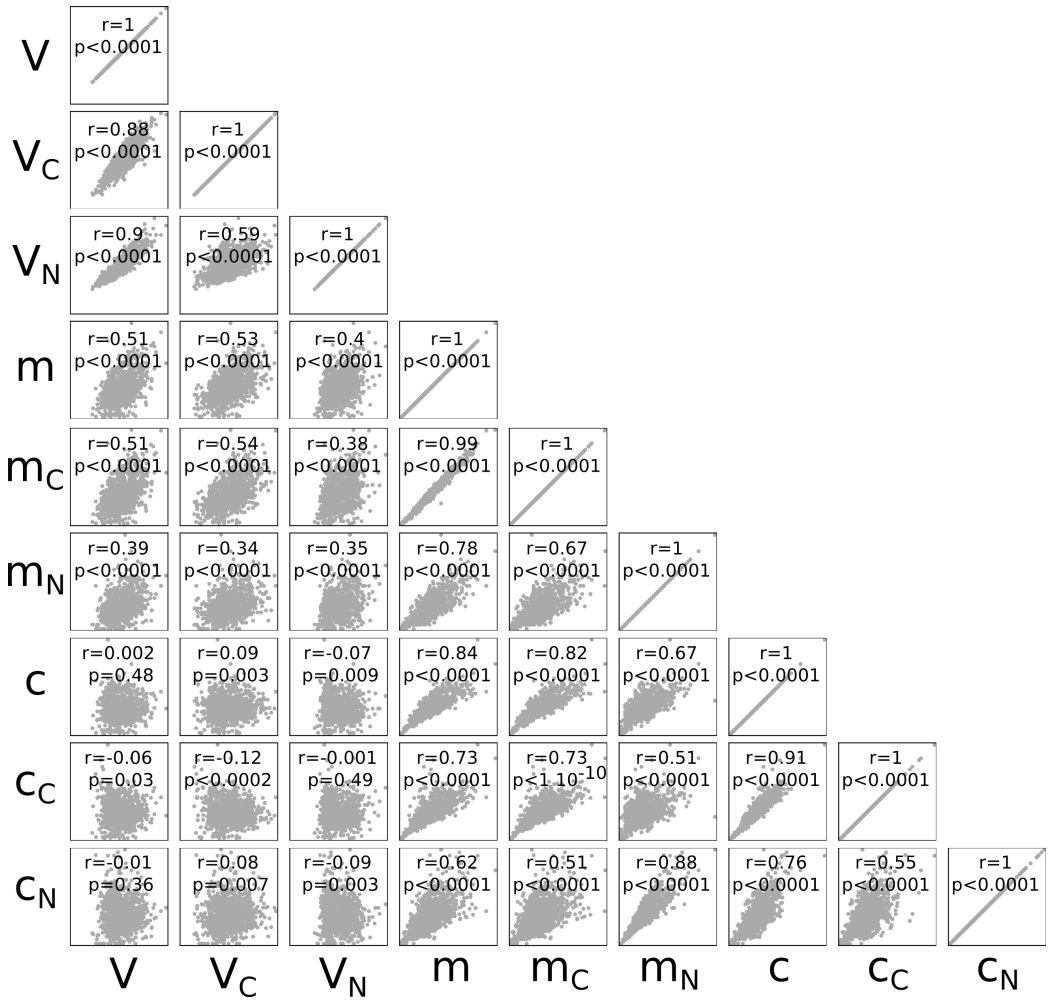

Figure 3.17: **Correlations within the data of Cell line II.** Within each plot the correlation coefficient(r) and the corresponding p-value are shown.

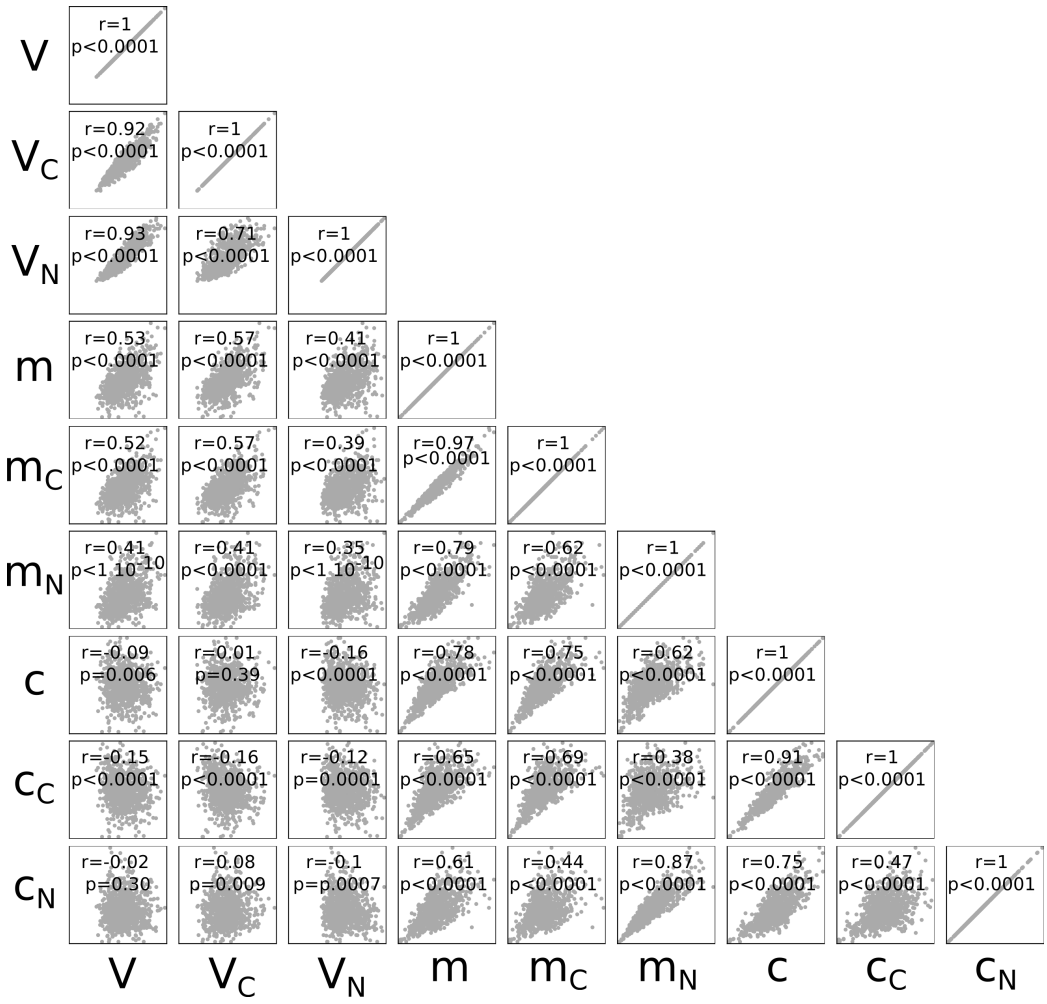

Figure 3.18: **Correlations within the data of Cell line III.** Within each plot the correlation coefficient(r) and the corresponding p-value are shown.

3.3.7 Probe sequence

The sequence of the mRNA FISH probe to detect the GFP reporter mRNA's.

	sequence	position	%GC
01	tcgcccttgctcaccat	1	59
02	accaccccggtgaacag	22	65
03	ccagctcgaccaggatg	42	65
04	gtggccgtttacgtcgc	62	65
05	tcgccggacacgctgaa	82	65
06	taggtggcatcgccctc	103	65
07	acttcagggtcagcttg	123	53
08	cttgccggtggtgcaga	143	65
09	agggtgggccagggcac	166	76
10	cgtaggtcagggtggtc	186	65
11	gcggctgaagcactgca	206	65
12	tgcttcatgtggtcggg	226	59
13	cggacttgaagaagtcg	246	53
14	gacgtagccttcgggca	266	65
15	aagaagatggtgcgctc	286	53
16	tgtagttgccgtcgtcc	306	59
17	aacttcacctcggcgcg	328	65
18	ttcaccagggtgtcgcc	349	65
19	tgcccttcagctcgatg	369	59
20	gccgtcctccttgaagt	389	59
21	agcttgtgccccaggat	409	59
22	ggctgttgtagttgtac	429	47
23	cggccatgatatagacg	450	53
24	gatgccgttcttctgct	470	53
25	ggcggatcttgaagttc	492	53
26	gctgccgtcctcgatgt	512	65
27	tagtggtcggcgagctg	532	65
28	cgatgggggtgttctgc	552	65
29	tgtcgggcagcagcacg	579	71
30	ctgggtgctcaggtagt	599	59
31	gggtctttgctcagggc	619	65
32	tgtgatcgcgcttctcg	639	59
33	cacgaactccagcagga	659	59
34	agagtgatcccggcggc	679	71
35	tgtacagctcgtccatg	699	53